

Target prediction for small, noncoding RNAs in bacteria

Brian Tjaden*, Sarah S. Goodwin¹, Jason A. Opdyke¹, Maude Guillier²,
Daniel X. Fu¹, Susan Gottesman² and Gisela Storz¹

Computer Science Department, Wellesley College, Wellesley, MA 02481, USA, ¹Cell Biology and Metabolism Branch, National Institute of Child Health and Human Development, Bethesda, MD 208902-5430, USA and ²Laboratory of Molecular Biology, National Cancer Institute, Bethesda, MD 20892, USA

Received February 2, 2006; Revised April 5, 2006; Accepted April 20, 2006

ABSTRACT

Many small, noncoding RNAs in bacteria act as post-transcriptional regulators by basepairing with target mRNAs. While the number of characterized small RNAs (sRNAs) has steadily increased, only a limited number of the corresponding mRNA targets have been identified. Here we present a program, *TargetRNA*, that predicts the targets of these bacterial RNA regulators. The program was evaluated by assessing whether previously known targets could be identified. The program was then used to predict targets for the *Escherichia coli* RNAs RyhB, OmrA, OmrB and OxyS, and the predictions were compared with changes in whole genome expression patterns observed upon expression of the sRNAs. Our results show that *TargetRNA* is a useful tool for finding mRNA targets of sRNAs, although its rate of success varies between sRNAs.

INTRODUCTION

In recent years, hundreds of RNAs that do not encode proteins but have intrinsic functions as regulators have been identified. These RNAs are generally denoted noncoding RNAs in eukaryotes and small RNAs (sRNAs) in bacteria. In *Escherichia coli* alone, >70 sRNA genes have been identified. Those bacterial sRNAs whose functions have been characterized can be sorted into three general categories: sRNAs that have intrinsic catalytic activity or are components of ribonucleoproteins, sRNAs that affect protein activity by structurally mimicking other nucleic acids and sRNAs that post-transcriptionally regulate mRNAs via basepairing interactions [reviewed in Refs (1,2)]. sRNAs in the

latter category appear to be the most abundant in *E.coli* (more than a third of the known sRNAs) and are the focus of our study.

All of the *E.coli* sRNAs that act by basepairing affect either the stability or translation of the mRNA target; in most cases the mRNAs are encoded in *trans* at positions on the chromosome distant from the sRNA. An example of a potential basepairing interaction that can lead to mRNA degradation is shown in Figure 1 for the RyhB sRNA and its target, the *sodB* mRNA. For all of the basepairing sRNAs that are *trans*-encoded, the basepairing interaction is interrupted by gaps in the pairing. In addition, the sRNAs in this class bind to the RNA chaperone Hfq, which has been shown to facilitate the interaction between some of the more well-characterized sRNAs and their targets (3,4). When interacting with sRNAs, Hfq appears to bind preferentially to unstructured AU-rich regions, frequently between more structured loop regions of the RNA (3–5). Despite increased understanding of the physiological roles of the basepairing sRNAs, the targets for only a subset of these sRNAs are known. In addition, although many sRNAs are thought to regulate more than one mRNA transcript, frequently only a small number of targets have been identified for a given sRNA.

While the targets of basepairing sRNAs in bacteria have remained elusive, there has been better success in identifying targets of microRNAs (miRNAs) in eukaryotes. The function of miRNAs in modulating mRNA stability and translation in eukaryotes is analogous to the function of many of the basepairing sRNAs in bacteria. A number of computational approaches have been employed successfully for the prediction of miRNA targets in plants (6,7), flies (8–12) and mammals (13–16). However, while the consequences of eukaryotic miRNA and bacterial sRNA interactions with their targets are similar, there are a number of important differences between these two classes of noncoding RNAs

*To whom correspondence should be addressed. Tel: +1 781 283 3354; Fax: +1 781 283 3642; Email: btjaden@wellesley.edu
Present addresses:

Sarah S. Goodwin, University of California, San Francisco, CA 94143, USA
Daniel X. Fu, California Institute of Technology, Pasadena, CA 91126, USA

© The Author 2006. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

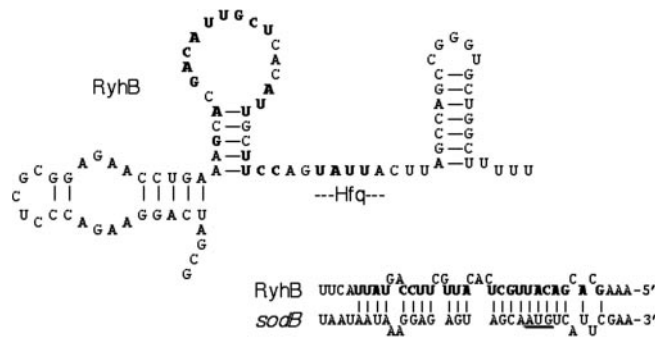


Figure 1. The figure depicts the secondary structure for the sRNA RyhB. The Sm-like protein Hfq binds to the AU-rich unstructured region of RyhB as indicated. Below the secondary structure, the primary sequence of RyhB is shown along with its putative binding interaction to the target mRNA *sodB* (42). The start codon for *sodB* is underlined. RyhB nucleotides that participate in the interaction are in bold.

which bring different challenges to the problem of target identification in bacteria. miRNAs are generally <25 nt in length, whereas sRNAs vary from ~50 to several hundred nucleotides in length. The shorter length of miRNAs helps focus the search for targets to more specific nucleotide sequences. The binding interactions between sRNAs and their targets also show more variation in the regions involved in pairing than in the case of miRNAs. For instance, in plants, miRNAs have a propensity to pair to mRNAs with near-perfect complementarity (6). In animals, target complementarity to the 5' portion of a miRNA (e.g. residues 2–8) may be critical for action (9,10,13). Also, in the case of miRNAs, target identification has been facilitated by restricting searches to particular regions of the target message, such as 3'-untranslated regions (3'-UTRs).

Here we present a program, *TargetRNA*, that can effectively predict mRNA targets of basepairing sRNAs. Several sRNAs with targets reported previously in the literature were tested with the program in order to validate the method. *TargetRNA* was then used to predict novel targets for a number of sRNAs. The results for four of the *E. coli* sRNAs, RyhB, OmrA, OmrB and OxyS, were investigated experimentally using northern and microarray analyses, leading to the identification of new targets for these sRNAs. Although only target predictions for *E. coli* sRNAs were experimentally tested, *TargetRNA* is also generally applicable to other bacteria. The program is publicly available at <http://snowwhite.wellesley.edu/targetRNA/>.

MATERIALS AND METHODS

Individual basepair model for hybridization scoring

The interaction between a given sRNA and a candidate mRNA target is predicted by calculating a hybridization score for the two RNA sequences. The individual basepair model of hybridization scoring is based on a straightforward extension of the Smith–Waterman dynamic program (17), except that instead of assessing homology potential, basepairing potential is assessed. Formally, let $S = s_1 s_2 \dots s_n$ be an sRNA sequence of n nucleotides and let $T = t_1 t_2 \dots t_m$ be a candidate target mRNA sequence of m nucleotides, where a subsequence $s_i s_{i+1} s_{i+2} \dots s_{j-1} s_j$ of S is denoted

as $S_{i,j}$ for any $1 \leq i \leq j \leq n$. The hybridization score h of two sequences S and T , with lengths n and m , respectively, is expressed recursively as follows:

$$h_{n,m} = \min \{ h_{n-1,m-1} + \delta(s_n, t_m), h_{n-1,m} + \Lambda_z, h_{n,m-1} + \Lambda_z, 0 \},$$

where $\delta(s_n, t_m)$ is the entry in matrix δ corresponding to the hybridization of nucleotide s_n with nucleotide t_m , and Λ_z is the score for a loop of length z in the interaction. Here, the 4×4 matrix δ represents the basepairing affinity of individual nucleotides, as opposed to the similarity of nucleotides as in the case of the Smith–Waterman algorithm. The default setting for the parameter δ is given by the matrix (A, C, G, U) \times (A, C, G, U) = [(6, 6, 6, -5), (6, 6, -5, 6), (6, -5, 6, 1), (-5, 6, 1, 6)], and the default setting for the parameter Λ_z is defined recursively as $3 + \Lambda_{z-1}$ if $z > 1$, and 12 if $z = 1$, following an affine score penalty for bulge and internal loops. Default parameter settings were determined by exploring the parameter space and evaluating the program's performance with a given set of parameters on the set of training data. Different parameter settings for δ and Λ_z did not yield significantly different results on the training set. The time requirement for this method is linear in the product of the RNA sequence lengths, $\Theta(nm)$.

Stacked basepair model for hybridization scoring

The stacked basepair model of hybridization scoring is based on stacking and destabilizing energies of interacting sequences. The calculation of the optimal hybridization score for two sequences using this model is comparable with the traditional approach for folding RNA sequences (18). The stacked basepair model calculates the minimum free energy of hybridization for two RNA sequences, without allowing intramolecular basepairings. Indeed, a number of RNA folding approaches such as MFold (19), the Vienna RNA Package (20), DINAMelt (21) and MultiRNAfold (22) enable estimation of the hybridization of two RNA sequences. Often these approaches work by concatenating the two sequences via a short linker sequence and then 'folding' the new concatenated sequence. The stacked basepair model is a straightforward extension of these approaches. Similar thermodynamic information and free energy parameters are used for loops and for stacked basepairs (23,24). Here, $\epsilon_{\text{stacked}}$ denotes the free energy parameter for a given pair of stacked bases, ϵ_{bulge} denotes the free energy parameter for a given bulge loop and its closing basepairs, and $\epsilon_{\text{internal}}$ denotes the free energy parameter for a given internal loop and its closing basepairs. Each of the free energy parameters may take a value of infinity if the closing nucleotides s_n and t_m do not basepair. The hybridization score h of two sequences S and T , with lengths n and m , respectively, is expressed recursively as follows:

$$h_{n,m} = \min \left\{ \begin{array}{l} h_{n-1,m-1} + \epsilon_{\text{stacked}}(s_{n-1}, s_n, t_{m-1}, t_m), \\ \min_{1 \leq i < n-1} \{ h_{i,m-1} + \epsilon_{\text{bulge}}(S_{i,n}, t_{m-1}, t_m) \}, \\ \min_{1 \leq j < m-1} \{ h_{n-1,j} + \epsilon_{\text{bulge}}(T_{j,m}, s_{n-1}, s_n) \}, \\ \min_{1 \leq i < n-1, 1 \leq j < m-1} \{ h_{i,j} + \epsilon_{\text{internal}}(S_{i,n}, T_{j,m}) \}, \\ 0 \end{array} \right\}.$$

Technically, the above formulation does not correctly reflect the thermodynamics of hybridization because it lacks energy contributions for dangling ends, terminal mismatches and initiation of hybridization. These omitted parameters are added after dynamic tabulation to appropriately reflect the free energy of hybridization of the two RNA sequences. The time requirement for this method is quadratic in the product of the sequence lengths, $\Theta(n^2m^2)$. However, if an upper bound is placed on the possible length of loops, then the time requirement is linear in the product of the sequence lengths, $\Theta(nm)$, though the hidden constant factors are much higher than in the case of the individual basepair model.

P-value calculation

Extreme-value distributions are well known to model the smallest (or largest) value among a set of independent random values. Let H be the optimal hybridization score determined by *TargetRNA* for two random RNA sequences. Then the distribution of H approximates an extreme-value type I distribution (25), whose probability density function is given by the following equation:

$$P(H = x) = \frac{1}{s} \exp\left(\frac{x-u}{s}\right) \exp\left(-\exp\left(\frac{x-u}{s}\right)\right),$$

where u is the location parameter and s is the scale parameter of the distribution (26). Accordingly, the cumulative distribution function is described by the following equation:

$$P(H > x) = 1 - \exp\left(-\exp\left(\frac{x-u}{s}\right)\right).$$

One of the program parameters which can be set before executing *TargetRNA* on a given sRNA gene is the searchable region of the candidate mRNA. For instance, a user of *TargetRNA* can choose to focus his or her search around the 5'-UTRs of messages as opposed to searching messages in their entirety. Searching longer regions of messages leads to lower expected hybridization scores, as illustrated by the distribution functions in Figure 2. To account for this dependency on the lengths of the sequences searched, following the use of Karlin-Altschul statistics (27), the hybridization score, h , is normalized by the log of the product of the sRNA sequence length, n , and the size of the mRNA search space, m , as follows:

$$h' = \frac{h}{\log(n * m)}.$$

Once a normalized hybridization score is computed, the P -value for the score can be calculated. In order to do so, however, the parameters u and s of the distribution of scores must first be determined. Ten thousand random RNA sequences are generated where the nucleotides of the random sequences are drawn from the first-order distribution of nucleotides contained in the actual mRNA search space. After computing normalized hybridization scores using the random sequences, the parameters u and s of the distribution of scores are estimated using the method of moments (28). With these parameter estimates, the probability of observing

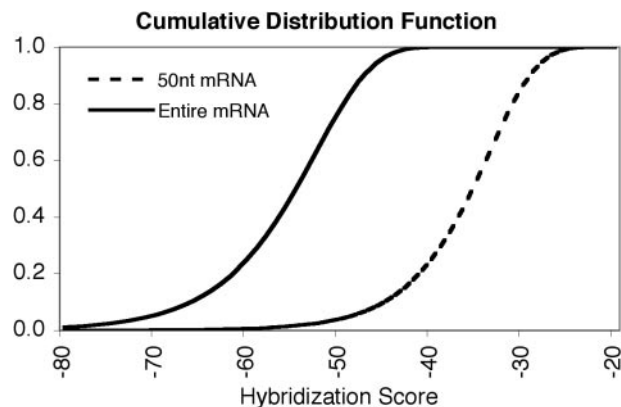


Figure 2. The graph shows the cumulative distribution function of hybridization scores for the sRNA RyhB. The curve on the right represents the hybridization scores calculated for RyhB against a message search space consisting of messages 50 nt in length (corresponding to a neighborhood around the ribosome-binding site). The curve on the left represents the hybridization scores calculated for RyhB against a message search space consisting of entire messages. Consequently, a hybridization score of -60 when searching messages 50 nt in length would lead to a significant P -value (<0.01), whereas a hybridization score of -60 when searching entire messages would not be significant.

a score equal to or less than h' by chance, i.e. the P -value, can be calculated as follows:

$$P(H \leq h') = \exp\left(-\exp\left(\frac{h' - u}{s}\right)\right).$$

Whole genome expression analysis

TargetRNA predictions were compared with whole genome expression data for four sRNAs. In each case, the sRNA was expressed from an inducible pBAD promoter in a strain deleted for the chromosomal copy of the sRNA gene, and the RNA levels were compared with those for an induced vector control strain. Duplicate experiments were performed; changes of 2-fold or better relative to the control were considered highly significant; changes of 1.5-fold or better were considered likely to be significant. Short expression times (15–20 min) were used to avoid some of the indirect effects of sRNA expression. MG1655 $\Delta ara714$ *ryhB::cat*/pBAD-RyhB was grown in Luria-Bertani (LB) medium to $A_{600} = 0.5$ and induced with 0.1% arabinose for 15 min (29). MG1655 $\Delta ara714$ $\Delta omrAB$ /pBAD-OmrA and MG1655 $\Delta ara714$ $\Delta omrAB$ /pBAD-OmrB were grown in LB medium overnight and induced with 0.2% arabinose for 20 min (30), and MG1655 $\Delta ara714$ $\Delta oxyS::kan$ (GSO112) carrying pBAD-OxyS was grown in LB medium to $A_{600} = 0.6$ and induced with arabinose for 15 min (J. A. Opdyke and G. Storz, unpublished data).

Strain construction and northern analysis

To assay the effect of OmrA and OmrB on *gntP* mRNA, it was necessary to first create a strain in which the *gntP* gene was expressed at a detectable level. This was done by creating a deletion of *uxuR*, which encodes a repressor of *gntP*, by homologous recombination. Briefly, the chloramphenicol resistance cassette was amplified with the Expand High Fidelity

PCR System (Roche) with oligonucleotides 5'uxuR::cm (GAT TAA CCG CAC CTA ACG GAC ACA ACA CCA TGA AAT CTG CCC CTG TGA CGG AAG ATC ACT TCG C) and 3'uxuR::cm (CGC AAG GAA CGT TTA CCC TTG CGC TTA TTA TAA TAA GTC AGG CTT ATC ACT TAT TCA GGC GTA GCA CC). The PCR product was then recombined into the chromosome of a DJ480 strain carrying a mini-lambda as described previously (31). The Δ uxuR::cm allele was then moved into the strain MG1099 (30) by P1 transduction to create strain MG1132.

MG1132 was transformed with pBR-plac-OmrA, pBR-plac-OmrB or the corresponding empty vector (30) and grown in LB medium with ampicillin to an A_{600} of 0.4, isopropyl- β -D-thiogalactopyranoside (IPTG) was added at a final concentration of 100 μ M, and the incubation was continued for 7 min. Total RNA was then extracted using hot phenol as described previously (32). Northern blot analysis for OmrA or OmrB was performed with 5 μ g total RNA, separated on an 8% urea-acrylamide gel (SequaGel; National Diagnostics) and transferred onto a positively charged nylon membrane at 200 mA for 1 h. For *gntP* and *ompA* (used as a loading control), total RNA (20 μ g) was separated on a 1% denaturing agarose gel and transferred onto a positively charged nylon membrane as in Ref. (33). Membranes were hybridized overnight at 42°C in Ultrahyb solution (Ambion, Austin, TX) with 100 ng/ml specific biotinylated probes. Detection was performed with the BrightStar BioDetect kit (Ambion) following manufacturer's instructions. Probes for OmrA, OmrB and *ompA* are as described in Ref. (30). The *gntP* probe was (Bio)-CTA CCG GTT GAT CTG CTT TCA GGA ATG ATG GCG TTG G.

To assay the effects of OxyS, overnight cultures of MG1655 Δ oxyS::cm (GSO113, generated by P1 transduction of the Δ oxyS::cm allele into MG1655) carrying pKK177-3 or poxyS (constitutively expressing OxyS) (34) were grown in LB medium with ampicillin to an A_{600} ~0.7. Total RNA was then isolated using hot phenol as above. Total RNA (5 μ g) from each strain was fractionated on a 1% denaturing agarose gel (33) (with or without 3.3% formaldehyde) together with a Millenium Marker (Ambion) and transferred to Zeta Probe GT membranes (Bio-Rad Laboratories,

Hercules, CA). The membranes were hybridized overnight at 45°C in Ultrahyb Oligo buffer (Ambion) with oligonucleotide probes (OxyS-A1: GCA GTG ACT TCA AGG GTT AAA AGA GGT GCC; yobF-1: GGC TCG GCA GAG AAG CGG TAT TCA ACG TCA ACG TG; wrbA-A1: TAA TTG CGG CGG CAT GGT TTC CGG TAC ACG; ybaY-1: CGG ATC CAG ACG GTA CCG GAG ACA TTC GGT TGC TGG) 5'-end labeled with 32 P using T4 polynucleotide kinase (New England Biolabs, Beverly, MA). The membranes were washed twice with a solution of 2 \times SSC and 0.1% SDS, first 30 s at room temperature and then 15 min at 45°C, and five times with a solution of 0.2 \times SSC and 0.1% SDS, each for 30 s at room temperature.

RESULTS

Training set

We first compiled a training set composed of putative mRNA targets of Hfq-binding sRNAs in *E. coli*, based on findings reported in the literature prior to 2005 (Table 1). The training set consists of 9 sRNAs interacting with a total of 12 message targets. For all but 2 of the 12 training examples (GcvB:*dppA* and GcvB:*oppA*), the putative location of interaction between the sRNA and its target mRNA has been described previously (Table 1).

These targets were then examined, along with their corresponding sRNA interactions, for common features. The binding interactions of the sRNAs with their mRNA targets contain gaps, mismatches and G:U basepairs. The longest stretches of contiguous nucleotides participating in duplex interaction range from 5 to 16 nt. In 8 of the 10 cases where the interaction has been described, the sRNA interacts with the message target near the translation start site (within ~30 bases of translation initiation). The two exceptions are DsrA:*rpoS* and RprA:*rpoS*, in which the target interaction occurs ~100 bases upstream of the translation start site and leads to positive rather than negative regulation of the *rpoS* mRNA (35,36). We also noted that, with the exception of DicF, which is processed from a longer transcript, each sRNA has a terminator stem-loop in its predicted structure. The OxyS:*fhlA* interaction is the only reported example of

Table 1. Putative mRNA targets of sRNA regulation in *E. coli* reported prior to 2005

sRNA	Target	Target function	Regulation	Target region of interaction (relative to AUG)	Reference	Score	P-value	Prediction ^a
DicF	<i>ftsZ</i>	GTPase involved in cell division	Negative	-28 to +2	(43)	—	—	—
DsrA	<i>hns</i>	Pleiotropic regulator	Negative	+7 to +19	(44)	-69	0.00098	#3
DsrA	<i>rpoS</i>	Sigma factor for stress response	Positive	-119 to -97	(35)	—	—	—
GcvB	<i>dppA</i>	Dipeptide transport protein	Negative	Unknown	(41)	-84	0.00014	#1
GcvB	<i>oppA</i>	Oligopeptide transport protein	Negative	Unknown	(41)	-70	0.00165	#4
MicC	<i>ompC</i>	Outer membrane pore protein	Negative	-41 to -15	(45)	-80	0.00021	#1
MicF	<i>ompF</i>	Outer membrane pore protein	Negative	-16 to +10	(46)	-80	0.00014	#2
OxyS	<i>fhlA</i>	Transcriptional activator	Negative	-15 to -9; +34 to +42	(37,39)	—	—	—
RprA	<i>rpoS</i>	Sigma factor for stress response	Positive	-117 to -94	(36)	—	—	—
RyhB	<i>sdh</i>	Succinate dehydrogenase	Negative	-42 to -3	(33)	-66	0.00215	#3
RyhB	<i>sodB</i>	Superoxide dismutase	Negative	-17 to +9	(42)	-60	0.00651	#9
Spot42	<i>galK</i>	Galactokinase in <i>gal</i> operon	Negative	-19 to +39	(47)	-78	0.00029	#1

^aFor each sRNA above, our computational approach was used to predict a set of candidate message targets of the sRNA. The final three columns in the table indicate the hybridization score of the predicted interaction, the P-value, and the rank (based on P-value) of the putative target among the set of predictions. For 4 of the 12 reported interactions, our approach did not predict the target with sufficient confidence (P-value < 0.01) using the default program parameters.

a terminator stem-loop participating in the target hybridization (37). Finally, with the exception of the DicF sRNA, each of the sRNAs shows evidence of conservation in closely related species to *E.coli* such as *Shigella flexneri* and *Salmonella typhimurium*.

Computational approach

We present a program, *TargetRNA*, which, given the sequence of an sRNA gene in a particular organism, outputs a ranked list of predicted message targets for the sRNA. The program begins by consulting a database of protein coding genes (38) for the organism of interest. For each protein coding gene in the organism, the program extracts the mRNA sequence corresponding to the protein coding region along with user-specified regions upstream and downstream of the coding sequence, extending into the 5'-UTR and 3'-UTR, respectively. *TargetRNA* then evaluates the potential for interaction between every extracted mRNA sequence and the sRNA, and assigns each a hybridization score and corresponding *P*-value (Materials and Methods). Finally, *TargetRNA* outputs a ranked list of the candidate message targets along with a graphical representation of each predicted interaction along the length of the sRNA. The program is freely available for use as a web application.

The interaction between a given sRNA and a candidate mRNA target is predicted by calculating a hybridization score for the two RNA sequences. In determining the hybridization score for two RNA sequences, intramolecular basepairings are not considered and pseudoknots are not allowed. To calculate the hybridization score of an sRNA and candidate mRNA target, *TargetRNA* can use either of two different hybridization score models for RNA sequence interactions: an individual basepair model or a stacked basepair model. The individual basepair model of hybridization scoring (described in Materials and Methods) is based on a straightforward extension of the Smith-Waterman dynamic program (17), except that instead of assessing homology potential, basepairing potential is assessed. The stacked basepair model of hybridization scoring (described in Materials and Methods) is based on stacking and destabilizing energies of interacting sequences, where the calculation of the optimal hybridization score for two sequences is comparable with folding RNA sequences (18) without allowing intramolecular basepairings.

Program parameters

In order to model the variations in action of individual sRNAs, *TargetRNA* has a number of user-adjustable parameters. For example, the program can use either of two different energy models, described above, for calculating the score of RNA:RNA hybridization. In addition, a seed, which corresponds to a minimum required length for at least one stretch of consecutive basepaired nucleotides in the RNA:RNA interaction, can be varied: different minimum numbers of contiguous nucleotides participating in the duplex interaction can be allowed and G:U basepairs may or may not be included. The seed is meant to reflect, biologically, the initial interaction between sRNA and mRNA, which has been shown in some cases to be a stretch of unpaired nucleotides in a loop of the sRNA that first basepairs with the target

message. Other program parameters include options for removing the terminator stem-loop of the sRNA from the hybridization calculation and restricting the search in the target mRNA sequence to a neighborhood around the ribosome-binding site, where most of the known interactions occur (Table 1). Finally, the threshold for the *P*-value of predicted hybridization interactions can be varied.

Evaluation of program parameters

To explore the effects of various parameter settings on the program's performance, the sensitivity and specificity of *TargetRNA* were evaluated with regard to the training set. The sensitivity (i.e. the true positive rate) is defined, for a given set of parameters, as the percentage of the 12 interactions in the training set which are correctly predicted by the program: True Positives/(True Positives + False Negatives). The specificity (i.e. the true negative rate) is defined, for a given set of parameters, as the percentage of non-interactions which are correctly predicted as non-interactions by the program: True Negative/(True Negatives + False Positives). For example, in *E.coli* each of the nine sRNAs in the training set may potentially interact with any of the 4244 messages. Thus, there are 9×4244 possible interactions, of which 12 are considered true interactions and the rest are considered non-interactions, for the purpose of evaluating the program's performance. In practice, the 9 sRNAs in the training set interact with >12 messages, so some of the program's predictions which we have classified as 'false positives' may indeed correspond to actual interactions. Thus, the estimated sensitivity and specificity of the program on the training set may in fact be conservative.

To evaluate the significance of target predictions, *P*-values rather than raw hybridization scores are employed. Hybridization scores have the disadvantage that they are dependent on the lengths of the sRNA and message sequences under consideration. As illustrated in Figure 2, longer sequences lead to lower expected hybridization scores than shorter sequences. Thus a hybridization score predicted for short sequences may be unlikely to have occurred by chance, whereas the same hybridization score for longer sequences may be likely to have occurred by chance. The *P*-value of a prediction corresponds to the probability of observing by chance a hybridization score at least as small as the predicted score. In other words, the *P*-value provides an indication as to the significance of a prediction. Based on evaluation of the program's performance on the training set, *P*-values ≤ 0.01 are considered significant.

Under all parameter settings tested, the individual basepair model of hybridization scoring (default method) resulted in a greater sensitivity on the training set than that of the stacked basepair model of hybridization (indicated as an option in the program). Closer inspection of the results revealed that the stacked basepair model favors longer regions of interaction between two RNAs, whereas the individual basepair model favors interactions of more parsimonious lengths. The bias of the stacked basepair model for longer basepairing interactions can be explained, in part, by the fact that the expected hybridization score for two random RNA sequences with the model is negative (i.e. favorable). Consequently, the stacked basepair model has a propensity for long, random interactions

as opposed to short, functionally meaningful interactions. Thus, this model may be more appropriate for identifying interactions in which the length of the interaction is known a priori, such as when the entire RNA gene participates in the interaction.

Using the individual basepair model, the performance of *TargetRNA* was assessed as each parameter was varied, in turn, while all others were held fixed. The receiver operating characteristic (ROC) curves in Figure 3 illustrate the trade-offs between sensitivity and specificity of the program on the training set as different parameters are varied. Each ROC curve is generated from 21 data points corresponding to the sensitivity and false positive rate ($1.0 - \text{specificity}$) as the seed is varied from 0 to 20 nt. The four different ROC curves demonstrate the performance of *TargetRNA* as the seed length varies when G:U basepairs are allowed in the seed interaction, when G:U pairs are disallowed in the seed, when the sRNA terminator loop is removed from the hybridization interaction, and when the sRNA terminator loop is retained. A similar analysis was carried out to identify the target sub-regions which yield the most advantageous sensitivity/specificity trade-off (data not shown).

The ROC analyses were then used to suggest default parameters for *TargetRNA*. Since the sensitivity, which ranges from 0 to 70%, appears to be more heavily influenced by the choice of parameters than the specificity, which ranges from 98.5 to 100%, default parameter values were chosen which minimize the false positive rate at the maximum sensitivity value, i.e. default parameters were chosen to correspond to the top left-most point along the ROC curves, as illustrated in Figure 3. Default parameters include removal of the terminator stem-loop of the sRNA, restricting the target message search from 30 nt upstream of translation initiation to 20 nt downstream of translation initiation, and necessitating a seed of at least 9 continuous nucleotides without G:U basepairs.

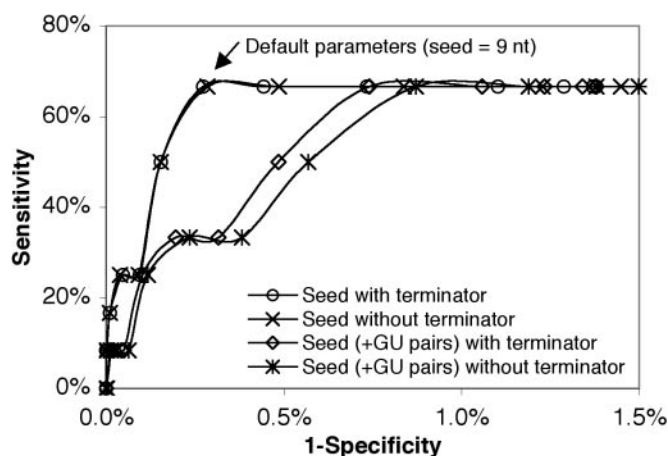


Figure 3. ROC curves depict the trade-offs in sensitivity (ordinate) and false positive rate (abscissa) on the training dataset. Each ROC curve is generated from 21 data points, as the initial seed of interaction is set to a value ranging from 0 to 20 nt. The four ROC curves correspond to different combinations of parameters, including allowing G:U wobble pairs in the hybridization seed, disallowing G:U pairs in the seed, removing the terminator loop from the hybridization score calculation and retaining the terminator loop. Default parameter settings correspond to the top left point among the ROC curves.

Performance on training set

Using the default parameters as described above, *TargetRNA* was run on the training set of 12 interactions between sRNAs and their targets. As illustrated in Table 1, for 8 of the 12 instances, the program predicted the reported message target among its set of top candidate targets. For each of these eight cases, the predicted interaction closely matched the reported interaction.

The four cases where the program did not predict the reported message target were examined more closely. In two of these four cases, namely *DsrA:rhoS* and *RprA:rhoS*, the reported interactions between the sRNAs and their targets occur ~100 nt upstream of the messages' translation start sites, outside of the region specified in our search. When the mRNA sequences were extended upstream to include >100 nt in the 5'-UTRs, *TargetRNA* predicted *DsrA:rhoS* as its top target candidate. *TargetRNA* did not predict the interaction between the sRNA OxyS and its target *fhlA*. The OxyS:*fhlA* interaction is the only example in the training set of a disjoint interaction where there are two separate regions of basepairing, which are >20 nt apart, the first region residing around the ribosome-binding site of the message target and the second residing downstream within the coding sequence. *TargetRNA* also did not predict the interaction between the sRNA DicF and its target *ftsZ*. The DicF:*ftsZ* interaction is the only example of an interaction where the longest stretch of contiguous nucleotides participating in the interaction is <7 nt. Altering the program parameters did not lead to prediction of either the OxyS:*fhlA* interaction or the DicF:*ftsZ* interaction. *TargetRNA*'s inability to predict, under any set of parameters, a few of the documented sRNA targets, suggests that the approach does not model effectively all sRNA:mRNA interactions. Given that 8 of the 12 targets were correctly predicted by the program, its sensitivity on the training set using the default parameters is ~67%. The specificity of the program on the training set using default parameters was estimated to be ~99% (Figure 3).

New predictions of *TargetRNA*

Given the paucity of sRNA targets which have been reported previously, as evinced by the small size of the training set, many of the predicted targets classified as false positives may actually be uncharacterized targets. To further explore how many of the high-scoring target candidates predicted by the program in fact correspond to novel message targets, we predicted targets for four sRNAs in *E.coli* (RyhB, OmrA, OmrB and OxyS) using somewhat more permissive parameters (removal of the terminator stem-loop of the sRNA, extending the target message search from 30 nt upstream of translation initiation to 30 nt downstream of translation initiation, and necessitating a seed of at least 7 continuous nucleotides with G:U basepairs, rather than a 9 nt seed). We then compared the output of *TargetRNA* with the results of whole genome expression analyses following induction of the sRNAs.

Although the four sRNAs chosen for the predictions are normally synthesized in response to different regulatory signals, for this work, we examined the effects of each of the sRNAs after brief expression from the ectopic pBAD promoter. For the RyhB, OmrA and OmrB sRNAs, it has been

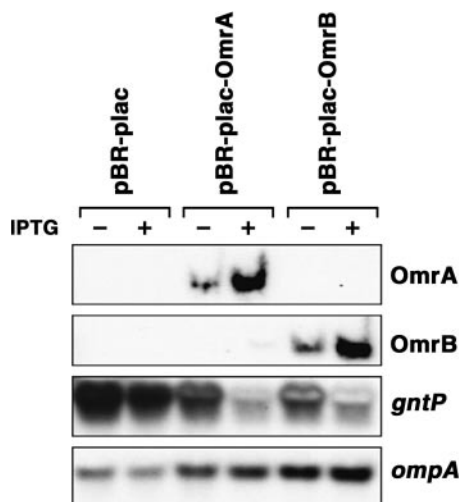


Figure 4. Northern analysis of the predicted OmrB target *gntP*. MG1132 (carrying the Δ *luxR::cm* allele) cells carrying pBR-plac, pBR-plac-OmrA or pBR-plac-OmrB were grown to midexponential phase ($A_{600} \sim 0.4$) and half of the sample was treated with IPTG for 7 min. Total RNA isolated from these samples was probed with oligonucleotides complementary to the indicated genes.

shown previously that negatively regulated target mRNAs are degraded upon pairing with an sRNA and this could be monitored using microarray analysis (29,30). We found mRNA targets of OxyS were also degraded and the effects likewise could be monitored by microarray analysis (see below, Figure 5). Consequently, the results of whole genome expression analyses were compared with the target predictions for each of the four above-mentioned sRNAs. It is worth noting two limiting assumptions in this approach. The first assumption is that the action of an sRNA on its targets will result in mRNA degradation. In previous work with RyhB and OmrA and OmrB, this was found to be the case for many targets, but may not be true in all cases; translational repression without degradation would lead to an underestimate of the specificity of *TargetRNA*. The second assumption is that the effects measured in the arrays are the direct effect of the sRNA, rather than secondary effects; this assumption might lead to an overestimate of the number of correctly identified targets. A relatively short expression time for the sRNAs was used to minimize this possibility. For the purposes of this study, changes in gene expression of at least 1.5-fold in duplicate experiments upon expression of the sRNAs were considered potentially significant, with changes of at least 2-fold considered highly significant. Genes for which the signals were deemed 'absent' or 'marginal' on the arrays for the vector control strains were excluded from the analysis. In some cases, northern blot analysis also was used to examine the levels of predicted target mRNAs. Although potential targets for which we observe a significant decrease in mRNA level are assumed to be direct targets of the sRNAs, further assays, such as compensatory mutations and/or *in vitro*-binding studies, are required to verify sRNA:mRNA basepairing.

RyhB target predictions. The 90 nt RyhB RNA is synthesized upon iron starvation and has been shown to cause the rapid

degradation of a number of target mRNAs upon pairing (29,32,33). Expression of RyhB from an ectopic promoter for 15 min, followed by examination of changes in mRNA abundance by microarray analysis, revealed potential targets, many of them corresponding to genes with the characteristics expected of RyhB targets, which primarily encode non-essential iron-binding proteins (29). RyhB is currently unique in the large number of known targets. From array analysis and other work, 18 operons and 56 genes have been shown to be regulated by RyhB, and interpreted as being directly regulated by this sRNA.

Among the predictions for RyhB from the *TargetRNA* program, 33 are below a *P*-value of 0.01; of these 15 were excluded from further analysis because the gene expression signals were too low to be significant (deemed as Absent, Table 2) or the gene was not on the array. Of the remaining 18 predictions 2 (*sdh* and *sodB*) were part of the training set; these 2 genes also show significant regulation in the arrays (6-fold and 19-fold decrease in signal after RyhB expression). Of the remaining 16 predictions, 2 others have changes in the array signal of >2-fold, and 6 others have signal changes between 1.5-fold and 2-fold (Table 2). Thus for RyhB, 10 out of 18 predictions (56%) are supported by the microarray data. It is worth noting that a number of targets for RyhB, thought to be direct targets based on either array or northern analysis, do not appear in the list of predicted targets (29,33); modification of program parameters leads to predictions for some but not all of these.

OmrA and OmrB target predictions. The 88 nt OmrA RNA and 82 nt OmrB RNA (previously RygA and RygB) are two partially homologous sRNAs, induced under high osmolarity conditions. They have been shown to regulate a number of genes encoding outer membrane proteins and surface structures (30). These sRNAs were each expressed ectopically and again the microarray results were compared with the predictions from *TargetRNA*. The results are strikingly different from the RyhB results. For OmrA, 36 targets are predicted with a *P*-value below 0.01, of which 10 were excluded from the analysis because the gene was missing or the signal for the vector controls was low or marginal (Table 3). Of the 26 remaining predictions, none showed differential expression in the array experiments of at least 1.5-fold, our cut-off value. One predicted target, *fecD*, is in an operon with four other genes; in previous work it was demonstrated that the *fecABCDE* operon is regulated by OmrA (30), and, in fact, the array data for other genes in the operon exceeds the 1.5-fold cut-off. If we consider *fecD* a correct prediction, 1 out of 26 predictions (4%) for OmrA is supported by the microarray data.

For OmrB, 18 targets are predicted with a *P*-value < 0.01, with 5 excluded because of low expression signals in control experiments (Table 4). Of the remaining 13, one, *gntP*, has differential expression of at least 2-fold in the array data. The effect of OmrB on *gntP* was confirmed by northern analysis (Figure 4). Interestingly, the northern analysis revealed that increased levels of OmrA expression also resulted in decreased *gntP* levels suggesting that *gntP* is also a target of OmrA, although this was not predicted by *TargetRNA* using the above parameters. For OmrB the microarray data supports 1 out of 13 predictions (8%). Thus, the percentages

Table 2. Predicted targets for RyhB

Rank	Gene	B#	Score	P-value	pBAD-RyhB (1) ^a	pBAD-RyhB (2) ^a	pBAD (1) ^a	pBAD (2) ^a	Ratio (avg) ^b	Other information ^c
1	<i>kdpA</i>	b0698	−83	0.00013						A/A; A/A
2	<i>citG</i>	b0613	−75	0.00055						A/A; A/A
3	<i>frdA</i>	b4154	−71	0.0011	337	493	1650	1865	4.4	P/P; P/P; operon agrees
4	<i>napF</i>	b2208	−68	0.0020	92	83	139	188	1.9	A/P; A/P; operon agrees
5	<i>yagJ</i>	b0276	−67	0.0023						A/A; A/A
6	<i>sugE</i>	b4148	−67	0.0023						A/A; A/A
7	<i>sdhD</i>	b0722	−66	0.0028	1666	1322	11388	6681	6.1	P/P; P/P; training set
8	<i>yhcF</i>	b3219	−66	0.0028						A/A; A/A
9	<i>sodA</i>	b3908	−66	0.0028	351	466	3412	2234	7.1 ^d	P/P; P/P; Fur target ^d
10	<i>motA</i>	b1890	−65	0.0034						A/A; A/A
11	<i>pinH</i>	b2648	−65	0.0034						A/A; A/A
12	<i>ygeZ/hyuA</i>	b2873	−65	0.0034						A/A; A/A
13	<i>ykgE</i>	b0306	−64	0.0040	333	241	437	339	1.4	P/P; P/P; operon agrees
14	<i>ydaN</i>	b1342	−64	0.0040	246	187	239	276	1.2	P/P; P/P
15	<i>ynfF</i>	b1588	−64	0.0040						A/A; A/A
16	<i>yiaM</i>	b3577	−64	0.0040						A/A; A/A
17	<i>cysE</i>	b3607	−64	0.0040	707	940	1122	1182	1.6	P/P; P/P
18	<i>yciS</i>	b1279	−63	0.0048	426	485	769	708	1.6	P/P; P/P
19	<i>yegK</i>	b2072	−63	0.0048						A/A; A/A
20	<i>acpS</i>	b2563	−63	0.0048	511	617	787	843	1.5	P/P; P/P
21	<i>ygiQ</i>	b4469	−63	0.0048						not on array
22	<i>ybjG</i>	b0841	−62	0.0058	1545	1182	1089	831	0.7	P/P; P/P
23	<i>yecD</i>	b1867	−62	0.0058	449	484	430	499	1	P/P; P/P
24	<i>metH</i>	b4019	−62	0.0058	269	277	384	389	1.4	M/P; A/P
25	<i>yadS</i>	b0157	−61	0.00698						A/A; A/A
26	<i>perM</i>	b2493	−61	0.00698						P/M; P/P
27	<i>metI</i>	b0198	−60	0.0083	615	778	624	529	0.8	P/P; P/M
28	<i>proA</i>	b0243	−60	0.0083	294	295	245	465	1.2	M/P; P/P
29	<i>yagT</i>	b0286	−60	0.0083						A/A; A/A
30	<i>nagZ</i>	b1107	−60	0.0083	555	381	713	714	1.6	A/P; A/P; operon agrees
31	<i>sodB</i>	b1656	−60	0.0083	198	265	4462	4285	19.3	M/P; M/P; training set
32	<i>ygiT</i>	b3021	−60	0.0083	100	121	91	140	1.1	P/P; P/P
33	<i>dadA</i>	b1189	−59	0.0099	1742	1587	3914	3088	2.1	P/P; P/P

^aStandard Affymetrix signal determined for indicated genes in two independent experiments.

^bAverage of ratios for pBAD control/pBAD-RyhB signal for the two experiments. For ratios >2, the predicted targets are highlighted in dark gray. For ratios 1.5–2, the predicted targets are highlighted in light gray.

^cSignals were determined to be A = absent, M = marginal or P = present by standard Affymetrix program and are listed in the following order: pBAD-RNA (1)/pBAD (1); pBAD-RNA (2)/pBAD(2). Only predicted targets for which the signal was scored as P for both of the two pBAD samples were considered.

^d*sodA* is regulated by the Fur repressor, but repression is more complete after induction of RyhB. In a parallel experiment, carried out with a *fur* mutant background, the ratio of mRNA level for *sodA* in the vector-containing cells and RyhB expressing cells was 1.7, still sufficiently high to be considered a direct predicted target (29).

of target predictions which are supported by the microarray experiments are much lower for these two related RNAs than for RyhB.

OxyS target predictions. The 109 nt OxyS RNA was one of the first sRNAs to be characterized. The expression of this RNA is strongly induced in response to oxidative stress and the RNA has been proposed to play a role in protecting cells against the damaging effects of elevated hydrogen peroxide concentrations (34). Although OxyS overexpression leads to a dramatic change in protein expression patterns and the sRNA has been proposed to regulate the expression of >40 genes, only one direct target, the *fhlA* mRNA, has been characterized (37,39). As discussed above however, this OxyS:*fhlA* basepairing interaction was unusual in many respects. To further explore OxyS targets, the OxyS RNA was expressed ectopically and the results of microarrays were compared with the predictions of *TargetRNA* (Table 5). Among the 23 OxyS target predictions with *P*-values <0.01, 6 were excluded because of low expression in the control samples or absence of the gene. Among the remaining 17, one showed between 1.5-fold and 1.9-fold

decreases in expression and four showed at least 2-fold decreases upon OxyS expression. The effects of OxyS on three of the strongly regulated genes, *yobF* (which is in an operon with *cspC*), *wrbA* and *ybaY* were also confirmed by northern analysis (Figure 5) (no transcript of the expected size could be detected for *yaiZ*). Overall, 5 out of 17 predictions (29%) are supported by the microarray data.

DISCUSSION

We present a computational method for predicting targets of sRNA genes in bacteria. While numerous *in silico* approaches have been proposed recently for identifying targets of miRNAs in eukaryotes (6–16), there has been a relative dearth of such approaches for sRNAs in bacteria. This lack of *in silico* methods may be due, in part, to the paucity of reported targets of sRNA regulation. To date, *E.coli* contains the best-studied set of sRNAs and targets. We compiled a list of 12 such targets in *E.coli*, all described in the literature prior to 2005. This set of targets was examined for common features, and a computational method was developed for predicting novel targets. The effectiveness of the approach was

Table 3. Predicted targets for OmrA

Rank	Gene	b#	Score	P-value	pBAD-OmrA (1) ^a	pBAD-OmrA (2) ^a	pBAD (1) ^a	pBAD (2) ^a	Ratio (avg) ^b	Other Information ^c
1	<i>yrfC/hofN</i>	b3394	-80	0.00018						A/A; M/M
2	<i>lit</i>	b1139	-74	0.00054	71	57	72	38	0.8	P/P; P/P
3	<i>narH</i>	b1225	-71	0.00093						A/A; A/A
4	<i>deoR</i>	b0840	-69	0.0013	364	344	278	275	0.8	P/P; P/P
5	<i>yzgL</i>	b3427	-68	0.0016						A/M; A/P
6	<i>yadL</i>	b0137	-67	0.0019						A/P; A/A
7	<i>gmhB</i>	b0200	-67	0.0019	571	543	497	465	0.9	P/P; P/P
8	<i>nadA</i>	b0750	-67	0.0019	150	110	157	114	1.0	P/P; P/P
9	<i>glcD</i>	b2979	-67	0.0019	486	372	354	280	0.7	P/P; P/P
10	<i>cheZ</i>	b1881	-66	0.0023						A/A; A/A
11	<i>clpB</i>	b2592	-66	0.0023	1356	1207	1669	1147	1.1	P/P; P/P
12	<i>uup</i>	b0949	-65	0.0028	463	520	513	382	0.9	P/P; P/P
13	<i>yccS</i>	b0960	-65	0.0028	143	103	114	99	0.9	P/P; P/P
14	<i>cydD</i>	b0887	-64	0.0034	213	213	227	178	1.0	P/P; P/P
15	<i>ydbC</i>	b1406	-64	0.0034	623	513	582	446	0.9	P/P; P/P
16	<i>hisM</i>	b2307	-64	0.0034	535	891	486	495	0.7	P/P; P/P
17	<i>malK</i>	b4035	-64	0.0034						A/A; A/A
18	<i>sufD</i>	b1681	-63	0.0041	1151	894	970	771	0.9	P/P; P/P
19	<i>folA</i>	b0048	-62	0.0049	771	441	1013	544	1.3	P/P; P/P
20	<i>yadD</i>	b0132	-62	0.0049	102	184	115	153	1.0	P/P; P/P
21	<i>ybcS</i>	b0555	-62	0.0049	206	121	118	204	1.1	P/P; P/P
22	<i>mipA</i>	b1782	-62	0.0049	2121	2451	2097	2219	1.0	P/P; P/P
23	<i>ygjN</i>	b3083	-62	0.0049	52	84	76	115	1.4	P/P; P/P
24	<i>yhbE</i>	b3184	-62	0.0049	425	846	373	640	0.8	P/P; P/P
25	<i>xylH</i>	b3568	-62	0.0049	464	404	411	600	1.2	P/P; P/P
26	<i>ssuC</i>	b0934	-61	0.0059						A/A; A/A
27	<i>csiE</i>	b2535	-61	0.0059	488	639	607	583	1.1	P/P; P/P
28	<i>yaeP</i>	b4406	-61	0.0059						not on array
29	<i>hokD/relF</i>	b1562	-60	0.0070	162	81	203	122	1.4	P/P; A/P
30	<i>fdoI</i>	b3892	-60	0.0070	1947	2014	2216	1750	1.0	P/P; P/P
31	<i>fecD</i>	b4288	-60	0.0070	224	334	325	233	1.1 ^d	P/P; P/P: operon >1.5
32	<i>hokB</i>	b4428	-60	0.0070						Not on array
33	<i>ydhT</i>	b1669	-59	0.0084						A/A; A/A
34	<i>yeaZ</i>	b1807	-59	0.0084	591	564	543	483	0.9	P/P; P/P
35	<i>yfbT</i>	b2293	-59	0.0084	641	492	797	657	1.3	P/P; P/P
36	<i>rumA</i>	b2785	-59	0.0084	466	517	562	624	1.2	P/P; P/P

^{a,b,c}As defined in Table 2.^dOperon previously found to be regulated by OmrA (30).**Table 4.** Predicted targets for OmrB

Rank	Gene	b#	Score	P-value	pBAD-OmrB (1) ^a	pBAD-OmrB (2) ^a	pBAD (1) ^a	pBAD (2) ^a	Ratio (avg) ^b	Other information ^c
1	<i>yadL</i>	b0137	-68	0.0015						A/P; A/A
2	<i>yaiY</i>	b0379	-68	0.0015	83	92	73	51	0.7	P/P; P/P
3	<i>trxC</i>	b2582	-65	0.0025	450	509	506	574	1.1	P/P; P/P
4	<i>ypdD</i>	b2383	-63	0.0037						A/A; A/A
5	<i>yphD</i>	b2546	-63	0.0037	121	160	127	136	0.9	P/P; M/P
6	<i>gntP</i>	b4321	-62	0.0044	87	146	307	346	2.9	A/P; P/P; Figure4
7	<i>fldA</i>	b0684	-61	0.0053	2108	2328	1945	1275	0.7	P/P; P/P
8	<i>ybjE</i>	b0874	-61	0.0053	93	119	80	81	0.8	P/P; P/P
9	<i>srlB</i>	b2704	-61	0.0053	122	195	183	175	1.2	P/P; P/P
10	<i>ykfI</i>	b0245	-60	0.0064						A/A; A/A
11	<i>b2680</i>	b2680	-60	0.0064						A/A; A/A
12	<i>mutM</i>	b3635	-60	0.0064	258	342	243	255	0.8	P/P; P/P
13	<i>yjhI</i>	b4299	-60	0.0064	94	101	83	139	1.1	P/P; P/P
14	<i>yaaH</i>	b0010	-59	0.0077						A/A; P/A
15	<i>ybhT</i>	b0762	-59	0.0077	799	607	624	1060	1.3	P/P; P/P
16	<i>ykgJ</i>	b0288	-58	0.0093	85	116	93	137	1.1	P/P; P/P
17	<i>yeeE</i>	b2013	-58	0.0093	563	361	499	612	1.3	P/P; P/P
18	<i>yhdN</i>	b3293	-58	0.0093	658	754	736	649	1.0	P/P; P/P

^{a,b,c}As defined in Table 2.

evaluated on the training set of 12 reported targets as well as on sets of predictions for the RyhB, OmrA, OmrB and OxyS sRNAs, for which the predictions could be compared with results from whole genome expression analyses.

The percentage of computationally predicted targets for which there was experimental support from microarray and northern blot assays ranged from ~4 to 8% for the OmrA and OmrB RNAs up to 56% for the RyhB RNA. The different

Table 5. Predicted Targets for OxyS

Rank	Gene	b#	Score	P-value	pBAD-OxyS (1) ^a	pBAD-OxyS (2) ^a	pBAD (1) ^a	pBAD (2) ^a	Ratio (avg) ^b	Other information ^c
1	<i>yobF</i>	b1824	−85	0.00015	1605	1529	9319	10 372	6.3	P/P; P/P; Figure 5
2	<i>yfdH</i>	b2351	−83	0.00021	1179	979	1168	628	0.8	P/P; P/P
3	<i>yeaK</i>	b1787	−70	0.0021	439	443	559	312	1.0	P/P; P/P
4	<i>b0816</i>	b0816	−68	0.0029						A/A; A/A
5	<i>rpmG</i>	b3636	−68	0.0029	10 229	5188	8009	4381	0.8	P/P; P/P
6	<i>dppD</i>	b3541	−66	0.0041	571	1828	332	862	0.6	P/P; P/P
7	<i>fliE</i>	b1937	−65	0.0049						M/A; A/A
8	<i>sgcQ</i>	b4303	−65	0.0049						A/A; A/A
9	<i>yaiZ</i>	b0380	−64	0.0059	111	275	377	426	2.4	P/P; P/P
10	<i>ymcC</i>	b0986	−63	0.0070	174	138	132	156	1.0	M/P; P/P
11	<i>fabB</i>	b2323	−63	0.0070	8694	7503	8524	8488	1.1	P/P; P/P
12	<i>rumA</i>	b2785	−63	0.0070	342	327	619	501	1.7	P/P; P/P
13	<i>ybbB</i>	b0503	−62	0.0083	216	237	204	342	1.2	P/P; P/P
14	<i>yccE</i>	b1001	−62	0.0083	59	73	34	75	0.8	P/P; P/P
15	<i>wrbA</i>	b1004	−62	0.0083	138	165	377	622	3.3	P/P; P/P; Figure 5
16	<i>ydeO</i>	b1499	−62	0.0083						A/A; A/A
17	<i>yheN</i>	b3345	−62	0.0083	530	559	645	499	1.1	P/P; P/P
18	<i>pmbA</i>	b4235	−62	0.0083	434	480	494	541	1.1	P/P; P/P
19	<i>ygaR</i>	b4462	−62	0.0083						not on array
20	<i>ybaY</i>	b0453	−61	.0099	41	87	177	343	4.2	A/P; A/P; Figure 5
21	<i>moaD</i>	b0784	−61	0.0099	853	1189	931	1000	1.0	P/P; P/P
22	<i>yeaC</i>	b1777	−61	0.0099	2414	3071	3828	3004	1.3	P/P; P/P
23	<i>phnC</i>	b4106	−61	0.0099						A/A; A/A

^{a,b,c}As defined in Table 2.

success rates with different sRNAs may be due to a number of factors including limitations of the microarray analysis that lead us to mistakenly underestimate the success rate of the predictions and the possibility that the program is less useful for some RNAs than others.

The low success rate for RNAs such as OmrA and OmrB may be due to a number of caveats associated with our experimental analysis. If pairing of an sRNA frequently leads to translational inhibition without mRNA degradation, our assay method, which is dependent upon changes in the mRNA levels, would improperly count the result as negative. Future experiments that directly test translation will be necessary to address this possibility. In addition, because we only evaluated targets that were detected at a sufficient level in the vector control to be judged ‘present’, the nature of the targets and their level of expression may change our evaluation of success. For instance, it is possible that RyhB target mRNAs are more abundant, in general, than OmrA and OmrB target mRNAs. If this is the case, OmrA and OmrB targets would be more likely to be deemed ‘absent’ under the assayed growth conditions and the rate of success for our computational predictions could be underestimated.

A few caveats regarding the *TargetRNA* program also should be taken into consideration. The program does not account for the structures of either the sRNA or mRNA. It is possible that some predicted basepairing interactions do not occur because the corresponding regions of either the sRNA or the mRNA are occluded by secondary structure. In addition to structure, some other feature of either the sRNA or mRNA not accounted for by *TargetRNA*, such as the presence of an Hfq-binding site, may be required for productive basepairing. Finally, while all of the sRNAs examined here bind Hfq and are believed to act by basepairing, they may represent different classes of sRNAs and may not follow the same rules for basepairing. Because the

training set used to develop the *TargetRNA* program used RyhB and OxyS substrates, and not OmrA and OmrB substrates, the program is not optimized for the latter RNAs. As an attempt to address this issue, we revisited the program parameters with a new training set, derived from the experiments presented here and recent results from the literature. The new training set of 25 targets included a number of OmrA and OmrB targets. However, we did not find a set of parameters that led to significant improvement in recognizing targets for the OmrA and OmrB sRNAs.

Alternatively some sRNAs may act on only a limited number of targets, while others may have many targets. For RyhB and OxyS, 209 and 186 genes, respectively, showed at least 2-fold changes while for OmrA and OmrB, 34 and 24 genes, respectively, showed at least 2-fold changes after expression of each of the sRNAs in one microarray experiment. While some of the effects of RyhB expression are known to be indirect, there were still more global effects, in general, of RyhB expression than of OmrA or OmrB expression.

Despite some limitations of both the *TargetRNA* program and whole genome expression analysis, we suggest that the combination of the two approaches will be an effective approach for identifying direct targets for an uncharacterized sRNA. Functional annotation may also be a useful indicator for identifying candidate targets. In several cases, the set of targets predicted by *TargetRNA* for a given sRNA was enriched for genes that appear functionally similar. For instance, among the top candidate targets for the sRNA GcvB were mRNAs *gltI*, *livJ*, *livK*, *ytfT*, *aroP* and *argT*, all genes encoding periplasmic transport proteins. Similarly, a number of top candidate targets for the sRNA RyhB encode non-essential iron-binding proteins.

While the method was evaluated on targets of sRNAs in *E.coli*, the approach is applicable to bacteria more generally.

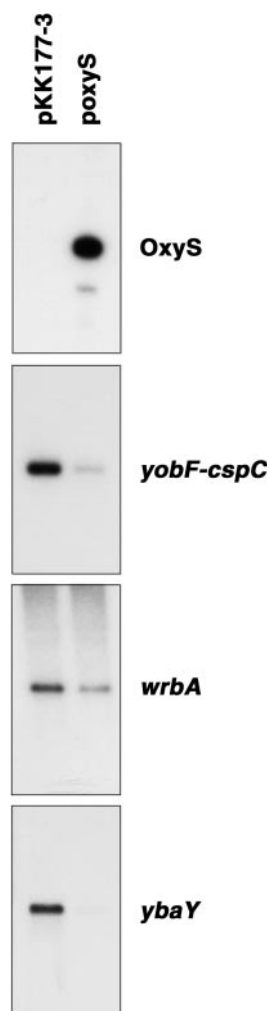


Figure 5. Northern analysis of predicted OxyS targets. MG1655 $\Delta oxyS::cm$ cells carrying pKK177-3 or poxyS were grown to late exponential phase ($A_{600} \sim 0.7$). Total RNA isolated from these samples was probed with oligonucleotides complementary to the indicated genes.

For example, in searching for targets of the sRNA BsrA in *Bacillus subtilis* (using a seed of at least eight nucleotides), the message target *rplU* which encodes a ribosomal protein was predicted. The *rplU* target of BsrA in *B. subtilis* has been documented previously (40). In addition, an ortholog of the *rplU* gene in *Listeria monocytogenes* was predicted as a target of the BsrA ortholog in *L. monocytogenes*. More generally, when searching for targets of an sRNA in a given organism, the program calculates the hybridization scores of orthologous targets with orthologous sRNAs in other bacteria. Since many sRNA genes are conserved across related species, the program can thus evaluate whether the targets are conserved and whether the hybridization interaction is conserved across species. One of the training examples was the GcvB:*dppA* interaction (41). For orthologous GcvB genes in *S. typhimurium*, *S. flexneri*, *Yersinia pestis* and *Photobacterium luminescens*, the program identifies orthologous *dppA* targets in all four bacteria. In each of the species, the hybridization score of the orthologous GcvB gene and the orthologous *dppA* target places the target among the top candidate predictions.

The application of *TargetRNA* to the *E. coli* RyhB, OmrA, OmrB and OxyS RNAs has already expanded the number of known targets for these regulatory sRNAs. We anticipate that as the number of known sRNA:mRNA interactions increases, we will better understand the applicability and the limitations of *in silico* target prediction approaches. In addition, an expanded set of known targets will allow for further refinements of computational approaches for target prediction.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank E. Massé for sharing his microarray results and P. Clote, D. FitzGerald, P. FitzGerald and E. Massé for comments on the manuscript. This work was supported by the Intramural Research Program of the NIH (NICHD and NCI). B.T. is supported by a Brachman Hoffman Fellowship at Wellesley College. Funding to pay the Open Access publication charges for this article was provided by Wellesley College.

Conflict of interest statement. None declared.

REFERENCES

- Gottesman, S. (2004) The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.*, **58**, 303–328.
- Storz, G. and Gottesman, S. (2006) Versatile roles of small RNA regulators in bacteria. In Gesteland, R.F., Cech, T.R. and Atkins, J.F. (eds), *The RNA World*, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 567–594.
- Zhang, A., Wassarman, K.M., Ortega, J., Steven, A.C. and Storz, G. (2002) The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs. *Mol. Cell*, **9**, 11–22.
- Møller, T., Franch, T., Hojrup, P., Keene, D.R., Bachinger, H.P., Brennan, R. and Valentin-Hansen, P. (2002) Hfq: a bacterial Sm-like protein that mediates RNA–RNA interaction. *Mol. Cell*, **9**, 23–30.
- Brescia, C.C., Mikulecky, P.J., Feig, A.L. and Sledjeski, D.D. (2003) Identification of the Hfq-binding site on DsrA RNA: Hfq binds without altering DsrA secondary structure. *RNA*, **9**, 33–43.
- Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B. and Bartel, D.P. (2002) Prediction of plant microRNA targets. *Cell*, **110**, 513–520.
- Wang, X.J., Reyes, J.L., Chua, N.H. and Gaasterland, T. (2004) Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biol.*, **5**, R65.
- Rehmsmeier, M., Steffen, P., Hochsmann, M. and Giegerich, R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.
- Stark, A., Brennecke, J., Russell, R.B. and Cohen, S.M. (2003) Identification of *Drosophila* microRNA targets. *PLoS Biol.*, **1**, 397–410.
- Lai, E.C., Tomancak, P., Williams, R.W. and Rubin, G.M. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol.*, **4**, R42.
- Robins, H., Li, Y. and Padgett, R.W. (2005) Incorporating structure to predict microRNA targets. *Proc. Natl Acad. Sci. USA*, **102**, 4006–4009.
- Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D.S. (2003) MicroRNA targets in *Drosophila*. *Genome Biol.*, **5**, R1.
- Lewis, B.P., Shih, I., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
- Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z. and Hatzigeorgiou, A. (2004) A combined

- computational-experimental approach predicts human microRNA targets. *Genes Dev.*, **18**, 1165–1178.
15. Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M. *et al.* (2005) Combinatorial microRNA target predictions. *Nature Genet.*, **37**, 495–500.
 16. John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C. and Marks, D.S. (2004) Human microRNA targets. *PLoS Biol.*, **2**, e363.
 17. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
 18. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
 19. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization predicting. *Nucleic Acids Res.*, **31**, 3406–3415.
 20. Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
 21. Markham, N.R. and Zuker, M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, **33**, W577–W581.
 22. Andronescu, M., Zhang, Z.C. and Condon, A. (2005) Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol.*, **345**, 987–1001.
 23. Xia, T., SantaLucia, J., Burkhard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C. and Turner, D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry (Mosc.)*, **37**, 14719–14735.
 24. Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 910–940.
 25. Altschul, S.F. and Gish, W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
 26. Gumbel, E.J. (1958) *Statistics of Extremes*. Columbia University Press, NY.
 27. Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
 28. Altschul, S.F. and Erickson, B.W. (1986) A nonlinear measure of subalignment similarity and its significance levels. *Bull. Math. Biol.*, **48**, 617–632.
 29. Massé, E., Vanderpool, C.K. and Gottesman, S. (2005) Effect of RyhB small RNA on global iron use in *Escherichia coli*. *J. Bacteriol.*, **187**, 6870–6873.
 30. Guillier, M. and Gottesman, S. (2006) Remodeling of the *Escherichia coli* outer membrane by two small regulatory RNAs. *Mol. Microbiol.*, **59**, 231–247.
 31. Yu, D., Ellis, H.M., Lee, E., Jenkins, N.A., Copeland, N.G. and Court, D.L. (2000) An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **97**, 5978–5983.
 32. Massé, E., Escorcia, F.E. and Gottesman, S. (2003) Coupled degradation of a small regulatory RNA and its mRNA targets in *Escherichia coli*. *Genes Dev.*, **17**, 2374–2383.
 33. Massé, E. and Gottesman, S. (2002) A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 4620–4625.
 34. Altuvia, S., Weinstein-Fischer, D., Zhang, A., Postow, L. and Storz, G. (1997) A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. *Cell*, **90**, 43–53.
 35. Majdalan, N., Cuning, C., Sledjeski, D., Elliott, T. and Gottesman, S. (1998) DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proc. Natl Acad. Sci. USA*, **95**, 12462–12467.
 36. Majdalan, N., Hernandez, D. and Gottesman, S. (2002) Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Mol. Microbiol.*, **46**, 813–826.
 37. Altuvia, S., Zhang, A., Argaman, L., Tiwari, A. and Storz, G. (1998) The *Escherichia coli* oxyS regulatory RNA represses *fhlA* translation by blocking ribosome binding. *EMBO J.*, **17**, 6069–6075.
 38. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomics, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
 39. Argaman, L. and Altuvia, S. (2000) *fhlA* repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense–target RNA complex. *J. Mol. Biol.*, **300**, 1101–1112.
 40. Suzuma, S., Asari, S., Bunai, K., Yoshino, K., Ando, Y., Kakeshita, H., Fujita, M., Nakamura, K. and Yamane, K. (2002) Identification and characterization of novel small RNAs in the *aspS-yrvM* intergenic region of the *Bacillus subtilis* genome. *Microbiology*, **148**, 2591–2598.
 41. Urbanowski, M.L., Stauffer, L.T. and Stauffer, G.V. (2000) The *gcvB* gene encodes a small untranslated RNA involved in expression of the dipeptide and oligopeptide transport systems in *Escherichia coli*. *Mol. Microbiol.*, **37**, 856–868.
 42. Geissmann, T.A. and Touati, D. (2004) Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J.*, **23**, 396–405.
 43. Tétart, F. and Bouché, J.P. (1992) Regulation of the expression of the cell-cycle gene *ftsZ* by DicF antisense RNA. Division does not require a fixed number of FtsZ molecules. *Mol. Microbiol.*, **6**, 615–620.
 44. Lease, R.A., Cusick, M. and Belfort, M. (1998) Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc. Natl Acad. Sci. USA*, **95**, 12456–12461.
 45. Chen, S., Zhang, A., Blyn, L.B. and Storz, G. (2004) MicC, a second small RNA regulator of Omp protein expression in *Escherichia coli*. *J. Bacteriol.*, **186**, 6689–6697.
 46. Andersen, J., Forst, S.A., Zhao, K., Inouye, M. and Delhas, N. (1989) The function of *micF* RNA. *micF* RNA is a major factor in the thermal regulation of OmpF protein in *Escherichia coli*. *J. Biol. Chem.*, **264**, 17961–17970.
 47. Möller, T., Franch, T., Udesen, C., Gerdes, K. and Valentin-Hansen, P. (2002) Spot 42 RNA mediates discoordinate expression of the *E. coli* galactose operon. *Genes Dev.*, **16**, 1696–1706.